Logical Modelling of Leslie's Theory Of Mind

Lee Flax

Computing Department, Macquarie University, Sydney
flax@ics.mq.edu.au

Abstract

Results of functional imaging of brain activity indicate that there is a brain component that handles "theory of mind" processes which lead to humans being able to work out the beliefs, desires, intetions and pretences of others. Leslie postulates a theory of mind that involves a decoupling process allowing a person to infer, for example, another's intentions etc. We model this theory of mind mechanism using first-order modal logic and give an example of a scenario involving pretence which uses the necessity and pretence modalities.

1. Introduction

Gallagher and Frith [7] write that humans have the ability to deceive others and enter into situations and games of pretence. It seems from this that humans are able to understand what is going on in another's mind. Even young children can do this. This capability is called having a "theory of mind". According to [1, page 161], having a theory of mind allows normal children to understand a range of mental states and to use them in making sense of and predicting action. They go on to say that at two years old children can understand pretending and the notion of desire; at three years old they understand that people have thoughts and understand things and at four years old they understand that people can have different and even false beliefs about the same state of affairs.

Aside from contributing to an explanation of the working of minds without disabilities, the theory of mind is also important in discussing the causes of disabilities such as autism [1, 6].

In this paper we will focus on only one aspect of the theory of mind: the cognitive behaviour involved in the activity of pretence and give a way of modelling it based on modal logic. In section 2 Leslie's model of the theory of mind is described. Then an informal account is given of logical modelling of the theory of mind in section 3. The syntax and semantics of modal logic is described in section 4, and it is applied to describe a scenario involving pretence in section 5.

2. Theory of Mind Mechanism

Gallagher and Frith relate in [7] that Leslie has given a model of the theory of mind and that he suggests that it "depends on a representation of imaginary circumstances 'decoupled' from reality" [8, 9]. Leslie in [8] postulates that a special cognitive brain component, which he calls the *decoupler*, is involved in handling the processing of pretence. In his model the decoupler has a two-way flow of information between itself and other central components of the cognitive system. The decoupler has three subfunctions: an expression raiser, a manipulator and an interpreter. The expression raiser isolates an expression to be involved in a pretence, say "this is a banana". The manipulator then proceeds with the pretence, "this banana is a telephone". The interpreter interacts with other central cognitive systems outside the decoupler to continue with the pretence.

In [7] Gallagher and Frith suggest on the basis of consistent results in several functional imaging studies of the human brain that the site of human mentalising ability (decoupling) is the anterior paracingulate cortex. This is located centrally towards the front of the head. Based on this suggestion it is postulated that it is possible to model the decoupling process of theory of mind by assuming that there is an actual cognitive component with a more or less precise location in the brain that is the site of the decoupler.

We abstract from the neuronal firing behaviour of brain components and suppose that the "language of thought" is first-order modal logic. This language is rich enough to represent thoughts expressed linguistically in a precise way and to also be able to deal with concepts such as pretence by making them corre-

Proc. 5th IEEE Int. Conf. on Cognitive Informatics (ICCI'06)
Y.Y. Yao, Z.Z. Shi, Y. Wang, and W. Kinsner (Eds.)
1-4244-0475-4/06/$20.00 ©2006 IEEE

spond to modalities. So, "Pretend the banana is a telephone", becomes a modal first-order logical sentence. In the next section we give an informal account of how this works logically and make it more precise in section 4.

3. Modelling Theory of Mind Mechanism

In this section we will set up semiformal machinery to model a simple scenario involving pretence. The treatment is kept semiformal here so that the arguments make intuitive sense and can be easily followed. A brief formal treatment will be given later in sections 4 and 5.

SCENARIO. We are playing with a child. A cup is held above a surface and we pretend that the cup holds liquid. When the cup is upturned we pretend that the surface is wet (even though it is not really wet).

The scenario given above is terse and relies heavily on one's understanding of the world to make sense. In this section we will tease out components of the semiformal modelling machinery to allow a start to be made on a logical treatment of the scenario.

We will need the following logical elements.

- Statements will be made using formal logical sentences. For example, a formal sentence will be used to systematise the statement, "A cup is held above a surface".
- We need a way of indicating in a formal way that certain sentences are assertions which apply everywhere. For example, a formal sentence will be used to systematise "It is necessarily so that if a cup containing liquid is upturned above a surface, then the surface will be wet". Such a formal sentence will be an example of the use of the *necessity modality*.
- We need a way of indicating in a formal way that a certain sentence is a pretence. For example, a formal sentence will be used to systematise "Pretend the cup contains liquid". Such a formal sentence will be an example of the use of the *pretence modality*.

The elements mentioned above are objects in an informal logical syntax. In addition to this syntax we need to be able to describe what these objects attach to or refer to in the world. This is semantics. Semantics describes the correspondence between syntactic objects and objects in the world. We will start by describing the informal syntax of our language and then we will begin to discuss the semantics. As result of our discussion of semantics we will see that our syntax needs to be treated more carefully and formally. This will be done in the next section and the semantics of our language will be given there as well.

The simplest sentences are formed from relation symbols. Relation symbols can have a finite number of blank places that can be filled with names of objects. An example of a relation symbol with one blank is: $\mathsf{contains\_liquid}(-)$. The fact that the Cup object contains liquid is written as: $\mathsf{contains\_liquid}(\mathsf{Cup})$. Another example of an object is $\mathsf{Surface}$. We use the relation $\mathsf{above}(-,-)$ with two blank places say that the cup is above the surface as follows: $\mathsf{above}(\mathsf{Cup},\mathsf{Surface})$. Complex sentences are formed from simpler sentences using symbols for propositional connectives, quantifiers and modalities.

The propositional connectives are negation $\neg$, conjunction $\wedge$, disjunction $\vee$, and implication $\rightarrow$. Here is how propositional connectives are used to form complex sentences from simpler ones. If X is a sentence then $\neg X$ is the sentence expressing $not X$. If X and Y are sentences then $X \wedge Y$ is the sentence expressing X *and* Y, $X \vee Y$ expresses X *or* Y, and $X \rightarrow Y$ expresses X *implies* Y or equivalently "If X, then Y".

The quantifier symbols are *there exists*, $\exists$, and *for all*, $\forall$. They are used to express facts about individual objects in the domain that the language addresses. For example, the sentence that there is an individual who is a man is written as $(\exists x)\mathsf{man}(x)$. The sentence that every individual is either a woman or not a woman is written as $(\forall x)(\mathsf{woman}(x) \vee \neg\mathsf{woman}(x))$.

The modalities we use are presented either in square or angle brackets. The shape of the brackets has to do with the way they are interpreted semantically. This will be described later (see definition 4.6). The *necessity* modality is written [Nec] and the the *pretence* modality <Pret>. The examples using modalities in the last two bullet points above translate as follows.

$$[\mathsf{Nec}](\mathsf{contains\_liquid}(\mathsf{Cup}) \wedge \mathsf{above}(\mathsf{Cup},\mathsf{Surface}) \wedge \mathsf{upturn}(\mathsf{Cup})) \rightarrow \mathsf{wet}(\mathsf{Surface}),$$

and

$$<\mathsf{Pret}>\ \mathsf{contains\_liquid}(\mathsf{Cup}).$$

To start the description of semantics, consider the plain sentence

$$\mathsf{contains\_liquid}(\mathsf{Cup})$$

(without the modal operator <Pret>). This sentence is true in the normal state of affairs in the real world if we first identify the relation symbol, $\mathsf{contains\_liquid}$, with the set of things that actually contain liquid and then check that the Cup object belongs to this set. Now consider the sentence <Pret> $\mathsf{contains\_liquid}(\mathsf{Cup})$. In

a pretend play scenario it is quite feasible that the Cup object is empty and yet we want to have a semantic correspondence which supposes the cup contains liquid. The key to enabling this is to allow a semantic set up which has several worlds. One can correspond to the real world and another to a world of pretence. Then in the real world the cup will correspond to an empty cup, but in the pretend world the cup corresponds to one holding liquid. So the syntactic Cup symbol corresponds to different cups in different worlds. Formal machinery needs to be used to allow the syntax and semantics to specify this accurately and unambiguously. First-order modal logic allows this to be done using *predicate abstracts*, see definition 4.1. Modal syntax and semantics are described in the next section.

4. First-order Modal Logic

We follow the treatment of first-order modal logic given by Fitting and Mendelsohn [3]. A good reference for propositional modal logic (that does not treat first-order logic) is Chellas [2].

Modal Syntax

In the informal treatment of syntax given in section 3 we allowed blank places of relation symbols to be filled with names of objects to give sentences, for example contains_liquid(Cup). In the development given here we do not fill in blank places with object names, instead we fill in blanks with with names of *variables* and then indicate how to substitute object names for the variables. This is done so that the semantics can be handled with precision, so avoiding difficulties caused when objects correspond to different things in different worlds. An example of this kind of difficulty was given at the end of section 3. The substitution process is carried out using *predicate abstracts*. These are described below.

Our modelling of the theory of mind mechanism is done with *sentences*. Sentences are special kinds of *formulas*, so we first describe how to build formulas. We do not need to use function symbols in our examples, so we will not include them in our syntax. However they are easily incorporated into the language, see [3, page 196]. We assume that we have the following.

- An infinite list of relation symbols. Each relation symbol has only finitely many places, one or more. We use letters such as p or words such as contains_liquid to denote relation symbols.
- An infinite list of variable symbols, denoted by x, y, x_1, etc.
- An infinite list of constant symbols, denoted by c, d, c_1, Cup, etc. These are used to name objects.
- Variable and constant symbols are called terms. So a term can be either a variable or a constant symbol. In our syntax terms take on this simple form because we do not use function symbols.

The definition of modal formula follows. In the definition we include the modality <Poss>, which denotes "It is possible that"; it is the dual of [Nec]. The notion of the free occurrences of variables in a formula is also defined because it is needed for the definition of a sentence, given below. (A sentence is a formula with no free variables.)

Definition 4.1 (Modal formula, sentence)

1. *Let p be an n-place relation symbol and $x_1, x_2, \ldots, x_n$ be variables, then $p(x_1, x_2, \ldots, x_n)$ is a formula. It is called an* atomic formula. *Every occurrence of a variable in an atomic formula is free.*
2. *If X is a formula, then so is $\neg X$. The free variable occurrences of $\neg X$ are the same as those of X.*
3. *If X and Y are formulas then so is $X \circ Y$, where $\circ$ is one of $\wedge$, $\vee$ or $\rightarrow$. The free variable occurrences of $X \circ Y$ are those of X together with those of Y.*
4. *If X is a formula, then so are [Nec]X, <Poss>X and <Pret>X. The free variable occurrences of each of these are the same as those of X.*
5. *If Y is a formula and x is a variable then $(\exists x)Y$ and $(\forall x)Y$ are formulas. The free variable occurrences of $(\exists x)Y$ and $(\forall x)Y$ are those of Y, but excluding occurrences of x.*
6. *If Y is a formula and x is a variable, then $(\lambda x.Y)$ is a* predicate abstract. *The free variable occurrences of $(\lambda x.Y)$ are those of Y, but excluding occurences of x.*
7. *If $(\lambda x.Y)$ is a predicate abstract and t is a term, then $(\lambda x.Y)(t)$ is a formula. The free variable occurrences of $(\lambda x.Y)(t)$ are those of $(\lambda x.Y)$ together with t if t is a variable. (We recall that a term is either a variable or a constant symbol.)*
8. *A* sentence *is a formula with no free variables.*

We now describe how to make syntactic items correspond to items in a world by setting up a semantics.

Semantics

Again we follow Fitting and Mendelsohn [3]. We saw earlier in section 3 that we needed to be able to describe things both in the real world and in a world of

pretence. So in our semantics we need more than one world. Modal semantics allows this very easily and naturally. We also need to be able to say how one world relates to another. This is done by specifying for each modality, a binary relation between worlds called the *accessibility relation*. The intuitive idea is that if two worlds are related, then one is accessible from the other, or one can see the other, or one is a modal alternative to the other. The worlds, their accessibility relations and the objects in each world are specified by defining a *frame*. For our modelling we only need a single accessibility relation to be used for all our modalities. If more were needed it would be easy to extend the frame definition to include more accessibility relations.

Definition 4.2 (Varying domain frame) *A varying domain frame $\mathcal{F}$ is a triple $(\mathcal{G}, \mathcal{R}, \mathcal{D})$, where*

1. *$\mathcal{G}$ is a set. Members of $\mathcal{G}$ are called worlds.*
2. *$\mathcal{R}$ is a binary relation on $\mathcal{G}$. $\mathcal{R}$ is called the* accessibility relation *on $\mathcal{G}$. If $(G, G') \in \mathcal{R}$ then G' is said to be accessible from G. The fact that $(G, G') \in \mathcal{R}$ is also written $G\mathcal{R}G'$.*
3. *$\mathcal{D}$ is a function mapping members of $\mathcal{G}$ to non-empty sets. For $G \in \mathcal{G}$, the set $\mathcal{D}(G)$ is called the* domain *of world G.*
4. *The union of all the world domains, $\bigcup\{\mathcal{D}(G) : G \in \mathcal{G}\}$, is called the* domain of the frame *$\mathcal{F}$ and is denoted $\mathcal{D}(\mathcal{F})$.*

We now specify what relation and constant symbols correspond to in a frame. This is done using an *interpretation*.

Definition 4.3 (Interpretation, model) *Let $\mathcal{F} = (\mathcal{G}, \mathcal{R}, \mathcal{D})$ be a frame. An* interpretation, *$\mathcal{S}$, defined on $\mathcal{F}$ is specified as follows.*

1. *If p is an n-place relation symbol and G is a world, then $\mathcal{S}$ assigns to (p, G) an n-place relation, denoted $\mathcal{S}(p, G)$, defined on $\mathcal{D}(\mathcal{F})$, the domain of the frame. So $\mathcal{S}(p, G)$ consists of n-tuples of elements from $\mathcal{D}(\mathcal{F})$.*
2. *If c is a constant symbol and G is a world, then $\mathcal{S}$ assigns to (c, G) an element, denoted $\mathcal{S}(c, G)$, of the domain of the frame, $\mathcal{D}(\mathcal{F})$. So $\mathcal{S}(c, G) \in \mathcal{D}(\mathcal{F})$.*
3. *If $\mathcal{S}$ is an interpretation defined on the frame $\mathcal{F}$, then the four-tuple $\mathcal{M} = (\mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{S})$ is called a* model.
4. *The* domain of the model *$\mathcal{M}$, denoted $\mathcal{D}(\mathcal{M})$, is the domain of its frame, $\mathcal{D}(\mathcal{F})$.*

A *valuation* is used to specify the correspondence between a variable and an element in a model's domain.

Definition 4.4 (Valuation, variant) *Let $\mathcal{M} = (\mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{S})$ be a model.*

1. *A* valuation *in $\mathcal{M}$ is a mapping v that assigns to each variable x an element, denoted $v(x)$, of the domain of the model $\mathcal{D}(\mathcal{M})$.*
2. *Let v and w be valuations in the model $\mathcal{M}$. The valuation w is an x-variant of v if v and w agree on all variables except possibly on the variable x.*
3. *The valuation w is an* x-variant of v at the world G *if w is an x-variant of v and $w(x) \in \mathcal{D}(G)$.*

Definition 4.5 (Term evaluation) *Let $\mathcal{M} = (\mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{S})$ be a model, G a world, v a valuation in $\mathcal{M}$ and t a term. We define $(v, \mathcal{S})(t, G)$ as follows.*

1. *If t is the variable x, then $(v, \mathcal{S})(t, G) = v(x)$.*
2. *If t is the constant symbol c, then $(v, \mathcal{S})(t, G) = \mathcal{S}(c, G)$.*

We are now able to define the truth of a formula in a model. The truth of a formula will depend on the model $\mathcal{M}$, the world G and the valuation v being used. If the formula X is true in the model $\mathcal{M}$ at the world G under the valuation v, we write $\mathcal{M}, G \Vdash_v X$.

Definition 4.6 (Truth in a model) *Let $\mathcal{M} = (\mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{S})$ be a model, G be a world in $\mathcal{G}$ and v a valuation in $\mathcal{D}(\mathcal{M})$ and t a term.*

1. *If p is an n-place relation symbol then $\mathcal{M}, G \Vdash_v p(x_1, x_2, \ldots, x_n)$ iff $(v(x_1), v(x_2), \ldots, v(x_n)) \in \mathcal{S}(p, G)$.*
2. *$\mathcal{M}, G \Vdash_v \neg Y$ iff $\mathcal{M}, G \nVdash_v Y$.*
3. *$\mathcal{M}, G \Vdash_v Y \wedge Z$ iff $\mathcal{M}, G \Vdash_v Y$ and $\mathcal{M}, G \Vdash_v Z$.*
4. *$\mathcal{M}, G \Vdash_v Y \vee Z$ iff $\mathcal{M}, G \Vdash_v Y$ or $\mathcal{M}, G \Vdash_v Z$.*
5. *$\mathcal{M}, G \Vdash_v Y \rightarrow Z$ iff $\mathcal{M}, G \nVdash_v Y$ or $\mathcal{M}, G \Vdash_v Z$.*
6. *$\mathcal{M}, G \Vdash_v (\forall x)Y$ iff for every x-variant w of v at G, $\mathcal{M}, G \Vdash_w Y$.*
7. *$\mathcal{M}, G \Vdash_v (\exists x)Y$ iff for some x-variant w of v at G, $\mathcal{M}, G \Vdash_w Y$.*
8. *$\mathcal{M}, G \Vdash_v [\mathsf{Nec}]Y$ iff for every $G' \in \mathcal{G}$, if $G\mathcal{R}G'$ then $\mathcal{M}, G' \Vdash_v Y$.*
9. *$\mathcal{M}, G \Vdash_v <\mathsf{Pos}> Y$ iff for some $G' \in \mathcal{G}$, $G\mathcal{R}G'$ and $\mathcal{M}, G' \Vdash_v Y$.*
10. *$\mathcal{M}, G \Vdash_v <\mathsf{Pret}> Y$ iff for some $G' \in \mathcal{G}$, $G\mathcal{R}G'$ and $\mathcal{M}, G' \Vdash_v Y$.*
11. *$\mathcal{M}, G \Vdash_v (\lambda x.Y)(t)$ iff $\mathcal{M}, G \Vdash_w Y$, where w is the x-variant of v such that $w(x) = (v, \mathcal{S})(t, G)$.*

Note that in general the truth conditions for <Pos> and <Pret> in the above definition will be different because their G' may be different.

5. Modal Account of the Scenario

We now give an account of the scenario using the syntax and semantics of modal logic. For convenience we repeat the scenario here.

SCENARIO. We are playing with a child. A cup is held above a surface and we pretend that the cup holds liquid. When the cup is upturned we pretend that the surface is wet (even though it is not really wet).

We translate the scenario into modal sentences and draw the conlusion that we can pretend that the surface is wet.

We need the following conditional statement to hold everywhere: if a cup containing liquid is above a surface and the cup is upturned then the surface is wet.

Let W stand for

$$(\mathsf{contains\_liquid}(x) \wedge \mathsf{above}(x, y) \wedge \mathsf{upturn}(x)) \rightarrow \mathsf{wet}(y),$$

then the following is supposed to hold everywhere

$$[\mathsf{Nec}](\lambda y.((\lambda x.W)(\mathsf{Cup})))(\mathsf{Surface}).$$

Let us pretend that the cup is above the surface, that it contains liquid and is then upturned. Let X stand for

$$(\mathsf{contains\_liquid}(x) \wedge \mathsf{above}(x, y) \wedge \mathsf{upturn}(x)),$$

then this is the statement of pretence

$$<\mathsf{Pret}> (\lambda y.((\lambda x.X)(\mathsf{Cup})))(\mathsf{Surface}).$$

We now argue semantically. Let $\mathcal{G} = (G, G')$, where G is the real world and G' is the world of pretence and suppose that $\mathcal{R} = (G, G')$. Suppose that $\mathcal{D}$ is specified and that the interpretation $\mathcal{S}$ is also specified; so the model $\mathcal{M} = (\mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{S})$ is specified. Let u be any valuation and suppose that

$$\mathcal{M}, G \vDash_u [\mathsf{Nec}](\lambda y.((\lambda x.W)(\mathsf{Cup})))(\mathsf{Surface}),$$

and

$$\mathcal{M}, G \vDash_u <\mathsf{Pret}> (\lambda y.((\lambda x.X)(\mathsf{Cup})))(\mathsf{Surface}).$$

Let v be the y-variant of u such that

$$v(y) = (u, \mathcal{S})(\mathsf{Surface}, G') = \mathcal{S}(\mathsf{Surface}, G')$$

and w the x-variant of v such that

$$w(x) = (v, \mathcal{S})(\mathsf{Cup}, G') = \mathcal{S}(\mathsf{Cup}, G').$$

From this it follows, using definition 4.6, that

$$\mathcal{M}, G' \vDash_w (\lambda y.\mathsf{wet}(y))(\mathsf{Surface})$$

and so

$$\mathcal{M}, G \vDash_w <\mathsf{Pret}> (\lambda y.\mathsf{wet}(y))(\mathsf{Surface}).$$

But

$$<\mathsf{Pret}> (\lambda y.\mathsf{wet}(y))(\mathsf{Surface})$$

is a sentence, it has no free variables, so using the result in Fitting and Mendelsohn [3, page 98]

$$\mathcal{M}, G \vDash_u <\mathsf{Pret}> (\lambda y.\mathsf{wet}(y))(\mathsf{Surface}),$$

which is the statement of pretence we wanted to prove.

6. Conclusion

Functional brain imaging studies have shown that there is a specific brain location involved in the human capability to fathom the beliefs, desires, pretences and intentions of others. This is called the theory of mind mechanism. We have modelled Leslie's theory of mind mechanism using modal possible world semantics, where we have postulated another world besides the real world in which the decoupling of pretense operates. Specifically, first-order modal logic has been used to model the pretence mechanism in the theory of mind. More work needs to be done to see whether logical modeling is robust enough to handle ongoing developments in this theory. An example of such a development is Tager–Flusberg's two component model of the theory of mind [10].

In another direction, the behaviour of cognitive components in inducing symptoms of schizophrenia has been modelled using a logic-based algebra [4]. This is based on models of cognitive functioning developed by Frith [5]. More work needs to be done to see how the algebraic and modal approaches can be integrated.

References

[1] Simon Baron-Cohen and John Swettenham. The relationship between sam and tomm: two hypotheses. In Peter Carruthers and Peter K. Smith, editors, *Theories of theories of mind*, pages 158–168. Cambridge university press, 1996.

[2] Brian F. Chellas. *Modal Logic.* Cambridge University Press, 1980.

[3] Melvin Fitting and Richard L. Mendelsohn. *First-order Modal Logic.* Kluwer Academic Publishers, 1998.

[4] Lee Flax. Algebraic modelling of some cognitive neuropsychology of schizophrenia. In Christine Chan, Witold Kinsner, Yingxu Wang, and D. Michael Miller, editors, *Third IEEE International Conference on Cognitive Informatics*, pages 131–137, Victoria, British Columbia, Canada, August 2004. IEEE.

[5] Christopher D. Frith. *The Cognitive Neuropsychology of Schizophrenia.* Lawrence Erlbaum Associates, 1992.

[6] Uta Frith. *Autism: explaining the enigma.* Blackwell, second edition, 2003.

[7] Helen L. Gallagher and Christopher D. Frith. Functional imaging of 'theory of mind'. *Trends in Cognitive Science*, 7(2):77–83, 2003.

[8] Alan M. Leslie. Pretense and representation: The origins of "theory of mind". *Psychological Review*, 94(4):412–426, 1987.

[9] Alan M. Leslie. Pretending and believing: issues in the theory of tomm. *Cognition*, 50:211–238, 1994.

[10] Helen Tager-Flusberg. What neurodevelopmental disorders can reveal about cognitive architecture, the example of theory of mind. In Peter Carruthers, Stephen Laurence, and Stephen Stich, editors, *The innate mind*, pages 272–288. Oxford university press, 2005.